FAST TRACK COMMUNICATION • OPEN ACCESS

Predator–prey survival pressure is sufficient to evolve swarming behaviors

To cite this article: Jianan Li *et al* 2023 *New J. Phys.* **25** 092001

View the article online for updates and enhancements.

New Journal of Physics
The open access journal at the forefront of physics

FAST TRACK COMMUNICATION

Predator–prey survival pressure is sufficient to evolve swarming behaviors

Jianan Li[1,2,*] ⓘ, Liang Li[3,4,5] ⓘ and Shiyu Zhao[2,6,*] ⓘ

1 Department of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, People's Republic of China
2 School of Engineering, Westlake University, Hangzhou 310024, People's Republic of China
3 Department of Collective Behavior, Max Planck Institute of Animal Behavior, 78464 Konstanz, Germany
4 Center for the Advanced Study of Collective Behavior, University of Konstanz, 78464 Konstanz, Germany
5 Department of Biology, University of Konstanz, 78464 Konstanz, Germany
6 Research Center for Industries of the Future, Westlake University, Hangzhou 310024, People's Republic of China
* Authors to whom any correspondence should be addressed.

E-mail: lijianan@westlake.edu.cn and zhaoshiyu@westlake.edu.cn

Abstract

The comprehension of how local interactions arise in global collective behavior is of utmost importance in both biological and physical research. Traditional agent-based models often rely on static rules that fail to capture the dynamic strategies of the biological world. Reinforcement learning (RL) has been proposed as a solution, but most previous methods adopt handcrafted reward functions that implicitly or explicitly encourage the emergence of swarming behaviors. In this study, we propose a minimal predator–prey coevolution framework based on mixed cooperative–competitive multiagent RL, and adopt a reward function that is solely based on the fundamental survival pressure, that is, prey receive a reward of -1 if caught by predators while predators receive a reward of $+1$. Surprisingly, our analysis of this approach reveals an unexpectedly rich diversity of emergent behaviors for both prey and predators, including flocking and swirling behaviors for prey, as well as dispersion tactics, confusion, and marginal predation phenomena for predators. Overall, our study provides novel insights into the collective behavior of organisms and highlights the potential applications in swarm robotics.

1. Introduction

Swarming behaviors in the nature, such as starling flocks, fish schools, and sheep herds [1], have been studied across various research fields, including biology [2], physics [3], and robotics [4]. Various agent-based models have been proposed to understand these collective behaviors. Most models are constructed based on designed static rules, including velocity alignment rules [3], a balance of social forces including attraction and repulsion [5, 6], and vision-based movement decisions [7–10]. Although these models can phenomenally give rise to swarming behaviors including complex ring-shaped and swirling patterns [6, 11, 12], the interaction rules are mostly heuristic and static, failing to capture the adaptation property of the biological world. To address this issue, recent studies have utilized evolution-based methods including neuroevolution [13] and reinforcement learning (RL) methods [14–17], because they offer the potential for adaptable strategies, analogous to the way biological organisms evolve [18–20]. In addition, multi-agent RL allows to model the interactions among individual agents in a swarm, and to optimize their collective behaviors [14–17].

However, the interaction mechanism or reward functions in evolution-based methods [13–17] are task-specific [21, 22] meaning they are handcrafted by designers to intentionally fit for the characteristics implicitly or explicitly associated to swarming behaviors, which we refer to as *swarm-dependent* in this study. For example, the reward function in [16] penalizes losing neighbors. That is, if an agent loses its neighbor, it receives a penalty of -1. The recent work in [17] assumes prey receive larger reward as the area of domain of

danger decreases [23], thus explicitly encouraging prey agents to move closer to each other. The authors in [13] implicitly assume the attack efficiency is inversely proportional to the number of prey visible to the predator due to confusion. As a result, the number of prey in low-density areas reduces faster than that in high-density areas, resulting in a clustering phenomenon as anticipated. A similar confusion mechanism is used in [15].

In this study, we propose a minimal predator–prey coevolution framework based on mixed cooperative-competitive multiagent RL, where the reward function is solely based on the motivation to survive for both predators and prey. Specifically, prey receive a reward of −1 if caught by predators while predators receive a reward of +1. This reward function has no relation to objectives such as increasing neighbors, decreasing sparsity, enhancing alignment, or promoting other characteristics directly associated with swarming behaviors, thus *swarm-independent*. Surprisingly, under the proposed framework, we find the simple survival pressure is sufficient to evolve flocking and swirling behaviors for prey. Quantitatively, we observe a steady increase in swarming density and group polarization. We also observe the emergent dispersion tactic, confusion and marginal predation phenomena for predators.

2. Modeling

2.1. Environment
We established a physics-based simulation environment where predators and prey interact with each other. The environment is a two-dimensional continuous space with two kinds of boundary conditions. As we will demonstrate in section 4, distinct boundary conditions will encourage flocking or swirling behaviors. The first kind is a finite square area, commonly used in previous studies such as [24, 25]. In this space, agents are unable to cross the boundaries, which are simulated as walls with a specified contact stiffness. The second kind is periodic boundary condition where an agent passes through one side of the square environment re-appears on the opposite side with the same velocity. This treatment wraps around the edges of the environment into a torus, thus enabling us to approximate a large or infinite space. Periodic boundary condition is widely used in molecular dynamics simulation [26] and swarm modeling [3, 16].

2.2. Agent dynamics
An agent, namely a predator or prey, is represented by a circle with a short line segment representing its heading, as shown in figure 1(a), where the unit vector $h \in \mathbb{R}^2, \|h\| = 1$ is the heading and $v \in \mathbb{R}^2$ is the velocity. The agents are subject to both active and passive forces.

The active force is self-generated propulsion to drive ego-motion, and consists of two components as shown in figure 1(a). The first component is the action to move forward aligned with the heading and computed as $a_F h \in \mathbb{R}^2$ where $a_F \in \mathbb{R}$. The second component is the action to rotate its heading, and is denoted as $a_R \in \mathbb{R}$ within a threshold value.

The passive forces include dragging force $f_d \in \mathbb{R}^2$, elastic force between contact agents $f_a \in \mathbb{R}^2$ as shown in figure 1(b), and elastic force between agents and boundaries $f_b \in \mathbb{R}^2$ as shown in figure 1(c). The dragging force is simply assumed to be in the opposite direction of velocity v with its magnitude proportional to $\|v\|$. Elastic forces f_a and f_b follow Hooke's law, and sum up when an agent contacts multiple other agents or boundaries as $f_a = \sum_j f_{a,j}$ and $f_b = \sum_j f_{b,j}$. The elastic forces prevent agents overlap and reflect physical collision dynamics. It is remarked that the velocity may not be aligned with the heading direction when a collision happens. In the simulation, the drag coefficient is set as $2\,\mathrm{N \cdot s\,m^{-1}}$ and the contact stiffness coefficient is set as $50\,\mathrm{N\,m^{-1}}$.

Combining the aforementioned active and passive forces, the dynamics of an agent of species i can be summarized as follows:

$$\dot{x} = v, \tag{1a}$$

$$\dot{v} = (a_F h + f_d + f_a + f_b)/m_i, \tag{1b}$$

$$\dot{\theta} = a_R, \tag{1c}$$

where $x \in \mathbb{R}^2$ is the position, $m_i \in \mathbb{R}^+$ is the mass, $\theta \in (-\pi, \pi] \subset \mathbb{R}$ is the heading angle and $h = [\cos\theta, \sin\theta]^\mathrm{T}$.
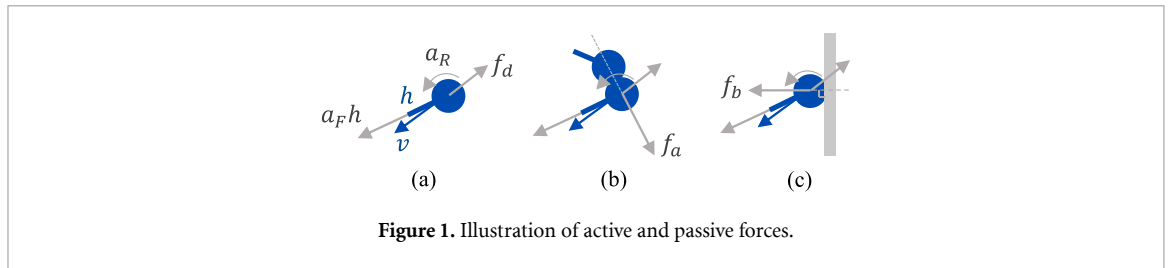
Figure 1. Illustration of active and passive forces.

2.3. Agent observation

The observation model of an agent is assumed to be dependent on both metric and topological distance. Metric dependence means an agent can only perceive others in its perception range [13, 14] which is assumed to be a disk with a pre-defined radius $R \in \mathbb{R}^+$ for simplicity. Topological dependence refers to how many at most an agent can perceive concurrently rather than how far away. The threshold is set as 6 which means an agent can perceive at most six allies and six adversaries. This setting is inspired by the work in [27] that each bird interacts on average with six neighbors, and further confirmed in the recent work [12]. The information contained in an observation vector includes relative position and heading of observed agents for both allies and adversaries, and the agent's own position, velocity and heading. If the number of agents in the perception range exceeds the topological threshold, the farthest ones are removed, and if it does not reach the threshold, the rest part of the observation vector are masked out with zeros. This topological modeling simplifies the structure of neural networks by fixing the input dimension.

3. Framework of coevolution

The optimization regime is set as a mixed cooperative–competitive multiagent RL framework where predators and prey are learning and adapting their behaviors concurrently analogous to the coevolution in the nature.

3.1. Homogeneity

Agents of the same species in a swarm are assumed to be homogeneous. This is reasonable when each agent has the same capability, responsibility, and goal as the others. More importantly, agents in a swarm typically display similar behavior patterns following the same interaction rule [7, 8, 28]. Homogeneity has been widely adopted in the modeling of swarm systems. For example, works such as those in [3, 5, 7–12] propose singleton control laws that dominate all the agents in a swarm, while those in [14–17] employ parameter-sharing techniques for the agents' neural networks. As a result, the homogeneity leads to two features of the proposed training regime: parameter-sharing of actor-critic networks and replay buffer sharing among conspecifics, which are described as follows.

3.2. Actor-critic

Conspecifics share one critic and one actor network. The critic network is used to evaluate the quality of an action taken by the agent, by estimating the expected future reward from that action, while the actor determines the agent's action $a = [a_{\mathrm{F}}, a_{\mathrm{R}}]^{\mathrm{T}}$ based on its current observation. Together, the critic and actor work to improve the agent's behavior over time, by continuously refining its estimates of expected future rewards and adjusting its policy accordingly.

The critic is designed to be decentralized allowing agents to evaluate based solely on local observations, without the knowledge of global states and actions like centralized critics used in [25, 29]. This is analogous to the situation when living organisms can only perceive nearby surroundings with limited abilities. Additionally, a decentralized critic provides better scalability as the number of agents in the system increases, making it particularly beneficial for swarm systems.

The actor is shared by conspecifics but not for their adversaries. Policy-sharing technique has been widely used in traditional agent-based models [3, 5, 7–12], and learning-based methods [14–17]. An advantage of policy-sharing is that the trained policy can be deployed across any number of agents of the same kind making it perfect for swarming research. It is remarked that although the parameters of the policy are shared, the output actions from the policy for each agent are different because they have different observations.

3.3. Replay buffer

When an agent takes an action in a given state, it receives a reward and transitions to a next state. The replay buffer is used to store such transitions of the agent interacting with the environment in the form of tuples:

state, action, reward and next state. The replay buffer serves to reduce correlations between consecutive experiences by exposing the agent to a more diverse set of experiences, thus helping to avoid over-fitting and stabilize the training process.

Two replay buffers $\mathcal{B}_0$ and $\mathcal{B}_1$ are used for predators and prey, respectively. There is no need to construct replay buffers for each individual because the homogeneity allows for interchangeable use of the experiences collected by conspecifics. Therefore, experiences collected from multiple conspecifics can be congregated into a single replay buffer, resulting in more resilient and effective learning outcomes.

3.4. Rewards

The reward for prey is simply set as $r = -1$ if it is caught by a predator, where the catch is represented by a contact between the two agents. Compared to the swarm-dependent interaction mechanisms or reward functions in [13–17], the proposed reward function solely emphasizes the potential dangers of remaining in close proximity to predators, thus swarm-independent. Similarly, the reward for a predator is $r = +1$ if it catches prey. It is remarked that prey agents are not removed from the simulation after being caught by predators. The contact between predators and prey can be likened to a continuous process where predators extract energy from the prey while engaged in 'eating' them. Upon separation, the prey's survival reward returns to zero, signifying the cessation of energy transfer analogous to the termination of 'bleeding'. A decorative reward that mimics energy consumption due to movement is added to the essential survival reward, which is simply set as $-0.01|a_\text{F}| - 0.1|a_\text{R}|$. This reward function will cause the agent to exhibit laziness.

In the special case when boundaries exist in the environment, an additional penalty -0.1 is added to the reward function when contact between agents and boundaries happens. This setting is designed to simulate either the presence of danger in the outside world or the situation where an agent chooses not to leave a specific location, such as a food-rich coral reef.

3.5. Algorithm

The algorithm primarily adopts the multiagent deep deterministic policy gradient method as proposed in [25, 30]. However, certain modifications have been made to the original method, which are summarized as follows: (1) to account for the limited perception capabilities of agents, a decentralized critic has been employed in lieu of a centralized one; (2) agents of conspecifics share one actor network but not for their adversaries; (3) experiences collected by conspecifics are merged into a single replay buffer for more effective learning.

We designate the predators and prey as species 0 and 1, respectively. The dimensionality of the observation and action vectors are denoted as d_o and d_a, respectively. We denote the discount factor, which determines the weight given to future rewards, as γ, and the soft update rate, which determines the speed at which a target network is updated towards a learning network, as τ. The proposed algorithm is summarized in algorithm 1.

Algorithm 1. Multiagent deep deterministic policy gradient algorithm for predator–prey coevolution.

`// i = 0` for predators, `i = 1` for prey
for species $i = 0$ to 1 **do**
    Randomly initialize actor μ_i parametrized by θ_i^μ and critic Q_i parametrized by θ_i^Q;
    Initialize target actor μ_i' and target critic Q_i', $\theta_i'^\mu \leftarrow \theta_i^\mu$, $\theta_i'^Q \leftarrow \theta_i^Q$;
end
for episode $= 1$ to M **do**
    Randomly spawn n_0 predators and n_1 prey; receive observations $o_i \in \mathbb{R}^{n_i \times d_\text{o}}$;
    for t $= 1$ to max-episode-length **do**
        for *all agents in species i*, select actions $a_i = \mu_{\theta_i}(o_i) + \mathcal{N}_t \in \mathbb{R}^{n_i \times d_\text{a}}$ where $\mathcal{N}_t$ is Gaussian noise;
        Execute actions a_i, receive reward $r_i \in \mathbb{R}^{n_i}$ and new observations $o_i' \in \mathbb{R}^{n_i \times d_\text{o}}$;
        Store $(o_i, a_i, r_i, o_i') \in (\mathbb{R}^{n_i \times d_\text{o}}, \mathbb{R}^{n_i \times d_\text{a}}, \mathbb{R}^{n_i}, \mathbb{R}^{n_i \times d_\text{o}})$ in replay buffer $\mathcal{B}_i$;
        $o_i \leftarrow o_i'$;
        Randomly sample a mini-batch of $S \in \mathbb{N}^+$ samples $(o_i^j, a_i^j, r_i^j, o_i'^j) \in (\mathbb{R}^{d_\text{o}}, \mathbb{R}^{d_\text{a}}, \mathbb{R}, \mathbb{R}^{d_\text{o}})$ from $\mathcal{B}_i$;
        Set $y_i^j = r_i^j + \gamma Q_i'(o_i'^j, a_i'^j)|_{a_i'^j = \mu_i'(o_i'^j)}$;
        Update critics by minimizing the loss $\mathcal{L}(\theta_i^Q) = \frac{1}{S} \sum_{j=1}^S \left(y_i^j - Q_i(o_i^j, a_i^j) \right)^2$;
        Update actors using sampled policy gradient $\nabla_{\theta_i^\mu} J \approx \frac{1}{S} \sum_{j=1}^S \nabla_{\theta_i^\mu} \mu_i(o_i^j) \nabla_{\mu_i(o_i^j)} Q_i(o_i^j, \mu_i(o_i^j))$;
        Soft-update target network parameters $\theta_i'^\mu \leftarrow \tau \theta_i^\mu + (1-\tau)\theta_i'^\mu$, $\theta_i'^Q \leftarrow \tau \theta_i^Q + (1-\tau)\theta_i'^Q$;
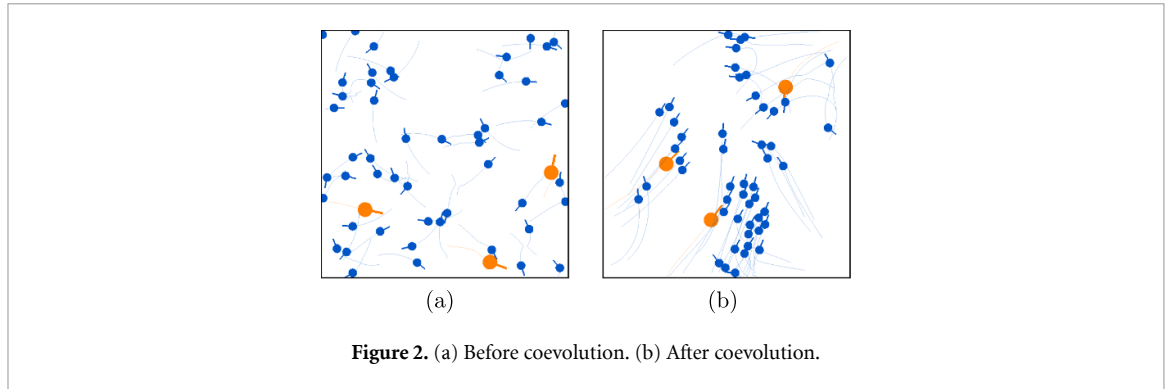   **end**
end

Figure 2. (a) Before coevolution. (b) After coevolution.

4. Results and analysis

4.1. Simulation setup

In the training phase, we instantiate $n_0 = 3$ predators and $n_1 = 10$ prey in the environment. This selection of population size is based on the assessment of hardware performance, as having more agents in the environment requires longer computation time. Nevertheless, the population size can be arbitrary as the policy is decentralized. In the evaluation phase, we deploy the trained policy on 50 prey, to yield a more distinct visual effect of swarming behaviors, whilst maintaining the same number of predators.

In order to provide a quantitative assessment of swarming behaviors, we introduce two measures: degree of sparsity (DoS) and degree of alignment (DoA). The DoS $\in [0, 1] \subset \mathbb{R}$ is defined as the average normalized distance to the nearest neighborhood of all conspecifics in an episode as

$$\text{DoS} = \frac{1}{TND} \sum_{t=1}^{T} \sum_{j=1}^{N} \|x_j(t) - x_k(t)\| \tag{2}$$

where $x_j(t)$ is jth agent position at time step t, $k = \arg\min \|x_j(t) - x_k(t)\|, k \in \{1, 2, \ldots, N\} \backslash j$, $T \in \mathbb{N}^+$ is episode length, $N \in \mathbb{N}^+$ is the total number which is equal to n_1 for prey, $D \in \mathbb{R}^+$ is the environment size defined as the maximum possible distance for two agents. For example, for a periodic square environment with edge length 2, the largest possible distance is $\sqrt{2}$. The symbol $\| \cdot \| : \mathbb{R}^2 \to \mathbb{R}$ denotes the Euclidean norm mapping. A smaller DoS indicates a denser swarm. An extreme case is when all conspecifics aggregate at the same point resulting in zero DoS. The definition of DoA $\in [0, 1] \subset \mathbb{R}$ is

$$\text{DoA} = \frac{1}{2TN} \sum_{t=1}^{T} \sum_{j=1}^{N} \|h_j(t) + h_k(t)\| \tag{3}$$

where $h_j \in \mathbb{R}^2$ is the heading of jth agent, k is the same as in the definition of DoS. It is worthy to remark that DoA is not equivalent to the mean heading of all conspecifics. This is because although the headings are similar in the same flock resulting in a high DoA, multiple flocks with different headings may cancel each other, as shown in figures 2(b) and 8(b). Therefore, calculating the relative quantity within a local neighborhood is more appropriate. A similar conclusion can be drawn for DoS.

4.2. Emergent flocking behavior

A typical scenario before and after evolution are shown in figures 2(a) and (b), respectively. The predators are depicted in orange, while the prey are in blue and a smaller size. Prior to evolution, the prey move randomly in different directions due to randomly initialized policy. Similarly, predators exhibit purposeless movements without the intention to pursue prey. The agents' behavior has undergone substantial changes after 2000 episodes of coevolution, as depicted in figure 2(b). Notably, the prey exhibit a remarkable emergence of cohesive movement patterns and a high DoA within multiple flocks, adhering to the well-known flocking basic rules outlined in [31].

The episodic evolution of running average of DoS and DoA is shown in figure 3, where the running average length is 100-episode and the shaded area indicates 95% confidence interval. Specifically, the DoS drops steadily from 22% of the environment size to around 19% suggesting a more cohesive movement of prey. Meanwhile, the DoA increases from 0.65 to approximately 0.82 indicating any two prey in the vicinity exhibit a higher DOA. It is remarked that the initial value of DoA is about 0.65. This is because for a uniformly distributed heading angle ranging from $-\pi$ to π, the expected DoA is given by
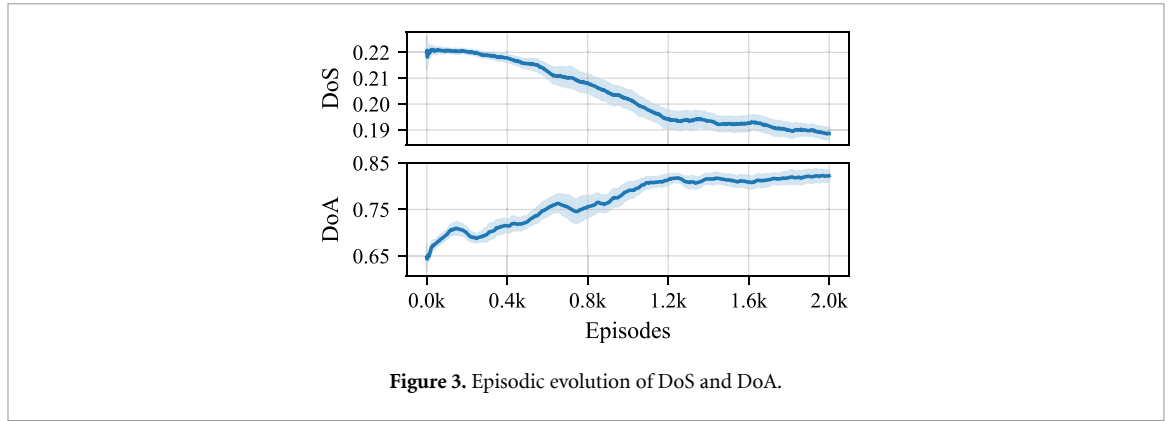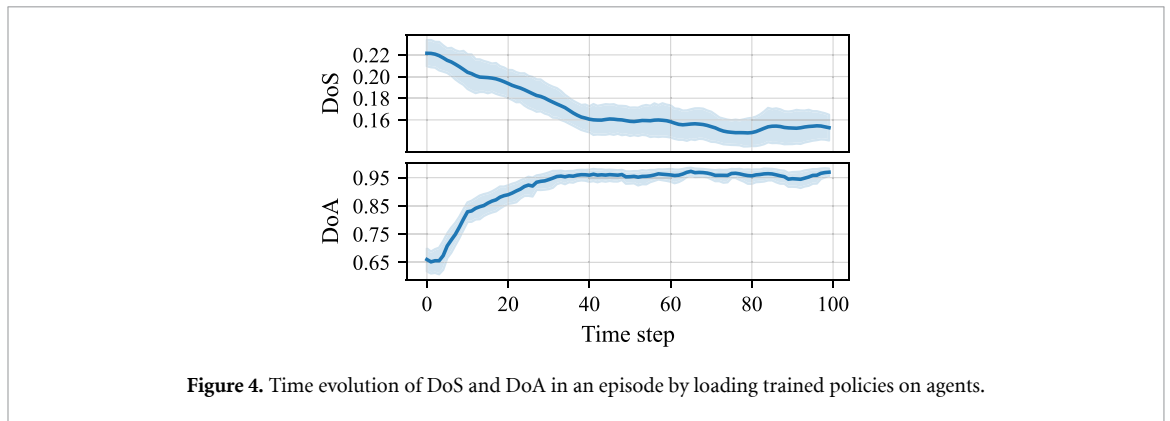
Figure 3. Episodic evolution of DoS and DoA.

Figure 4. Time evolution of DoS and DoA in an episode by loading trained policies on agents.

$\mathbb{E}[\cos(\phi/2)] = 2/\pi \approx 0.64$, where ϕ denotes the angle formed by two headings. This expected value is quite close to the value read from figure 3.

The DoS and DoA shown in figure 3 is a mean value computed from the entire episode, as indicated by $1/T \sum_{t=1}^{T}(\cdot)$ in equations (2) and (3). For established flocks, as shown in figure 2(b) or figure 7, the DoS is lower than 19% while its DoA exceeds 0.82. This can be shown by loading trained policies into agents, and plotting the DoS and DoA as functions of *time steps* in an episode, as shown in figure 4. In this case, the DoS decreases to approximately 15% while the DoA reaches around 0.96.

It is possible to question whether the observed alignment and distance reduction between neighboring prey is a result of swarming or merely a consequence of prey fleeing from predators in the same direction, which is commonly referred to as *herding*. To differentiate between these two phenomena, we conducted additional simulations where only trained prey agents are present. A typical scenario is illustrated in figure 5(a), where we observe that the prey still exhibit a high DOA even in the absence of predators, which is comparable to the well-known Vicsek model [3]. This finding further supports the hypothesis that swarming behavior observed in the absence of predators is a strategy employed by prey to avoid predators.

These findings are noteworthy because they reveal how simple predator–prey interaction, driven solely by the motivation to survive, could give rise to conspicuous flocking behavior that exhibits both cohesion and alignment characteristics. This leads us to hypothesize that the emergent flocking behavior is largely an outcome of passive space extrusion and polarization induced by predators.

4.3. Emergent confusion effect, dispersion tactic and edge effect

During the pursuit, predators exhibit confusion in certain situations, as illustrated in figures 6(a) and (b) at time step 30 and 33 of an episode. When the prey agent merges into a flock, the predator gives up the chase, slows down, and stagnates for a while as evidenced by a shorter path, appearing confused and uncertain about which prey to focus on.

Additionally, as shown in figure 7(a), predators often employ a *dispersion tactic*, in which they first move towards the center of a swarm to disperse it and then focus on isolated ones to capture [32, 33]. The resulting isolation or the phenomenon that predators frequently catch prey on the periphery of a swarm is referred to as *marginal predation* or *edge effect* [34, 35], as shown in figure 7(b). The observed marginal predation behavior suggests that predators may be less confused when targeting prey on the periphery of a swarm, potentially due to a smaller number of prey resulting in an elevated predation rate [33, 36], or a higher
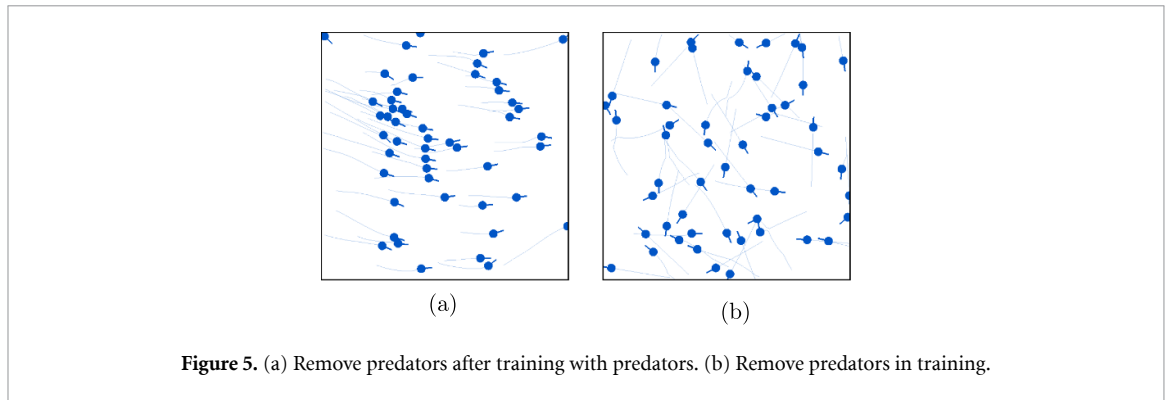
Figure 5. (a) Remove predators after training with predators. (b) Remove predators in training.
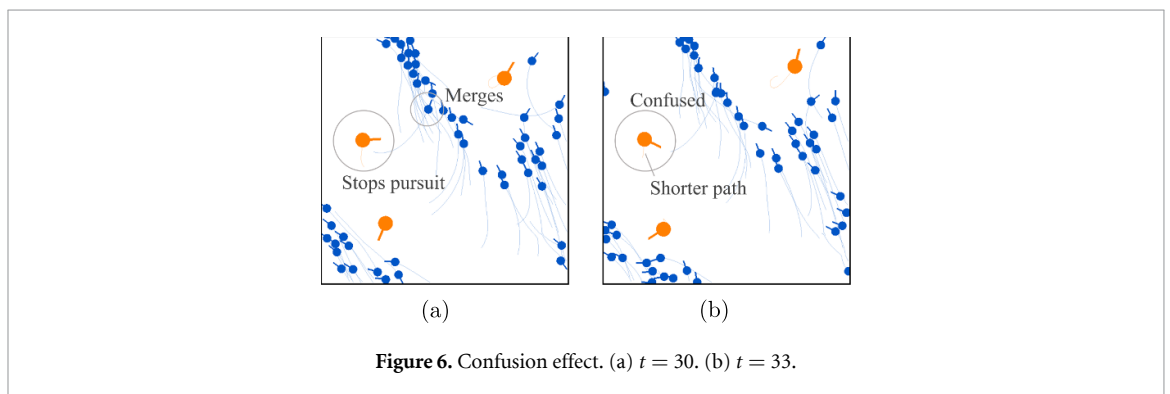

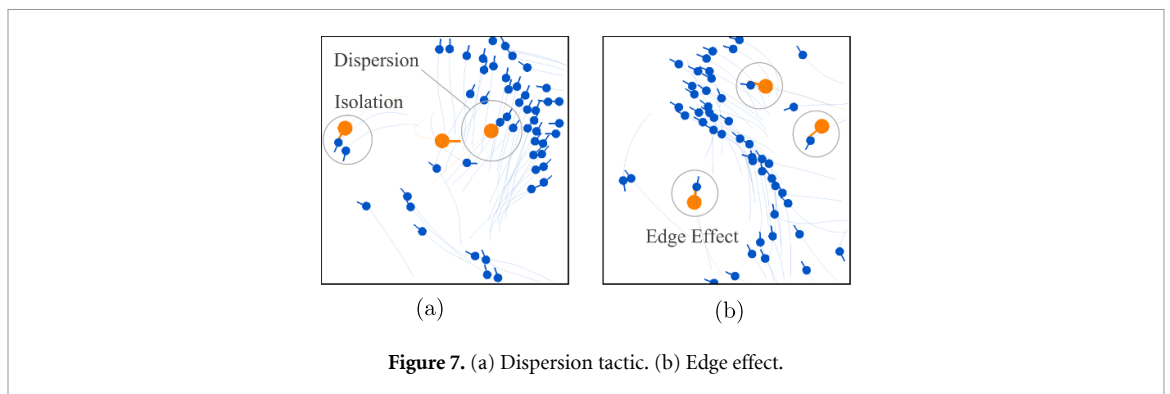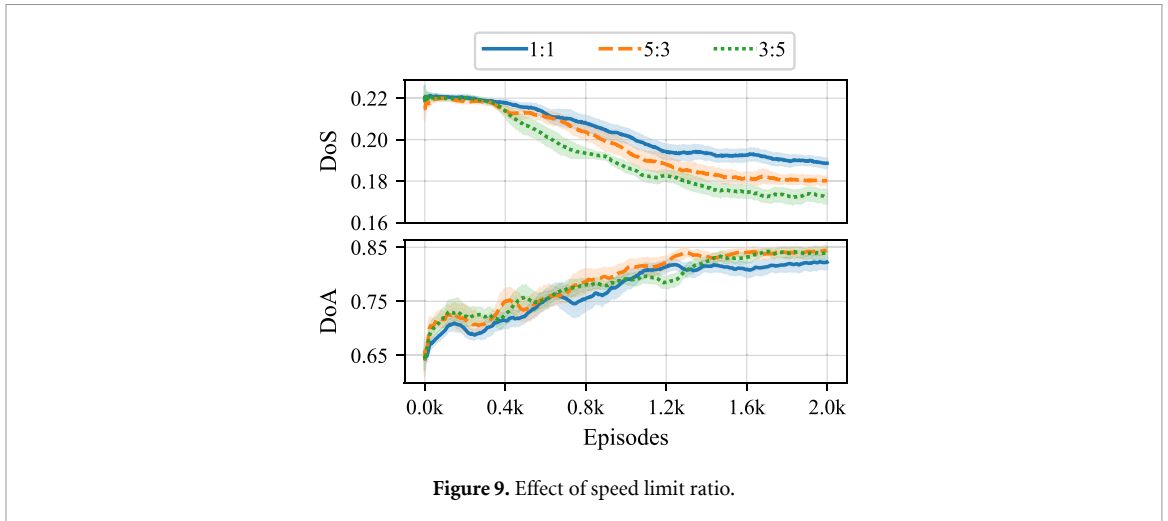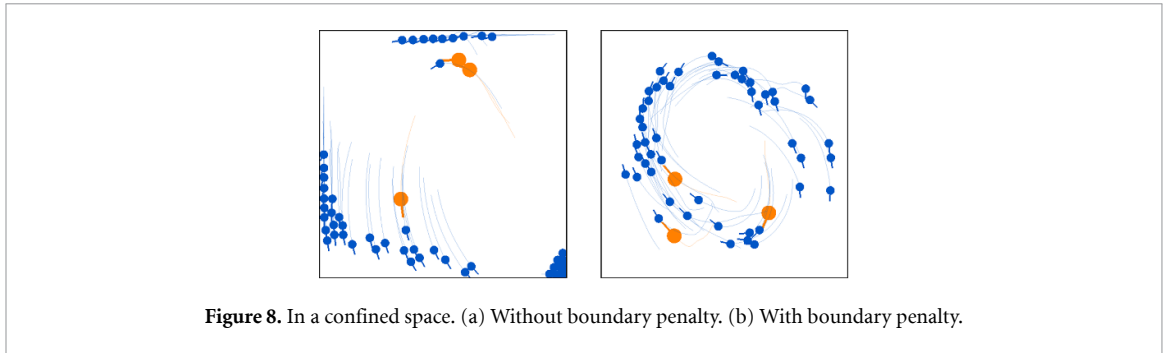**Figure 6.** Confusion effect. (a) $t = 30$. (b) $t = 33$.

Figure 7. (a) Dispersion tactic. (b) Edge effect.

encounter rate [34]. Overall, the emergent confusion effect and marginal predation behavior highlight the challenges predators face when trying to select the optimal target from a group of prey.

4.4. Emergent swirling behavior

We now examine the scenario when the environment is confined, and consider two cases: without and with penalties when prey collide with boundaries. It is remarked that the extra boundary penalty will complicate the environment, thus slowing down the learning process. To accelerate the evolution of prey, we endow the predators a behavioral rule that creates survival pressure by directly moving towards their nearest prey. This rule controls two active forces: first, the predator rotates its heading to point towards the nearest prey, and then it moves directly towards the target at maximum speed. By following this rule, the predators exert survival pressure on the prey and drive their evolutionary adaptation.

Without penalty, there are no significant differences in flocking behaviors compared to the infinite space case, except for boundary aggregation, as in figure 8(a), similar to fish swimming in a fishbowl or pond. With penalty, an enigmatic swirling behavior emerges as in figure 8(b). Such circular motion has been observed in fish and insects, yet it is hardly understood and still remains unclear in the scientific community [37, 38]. We hypothesize that the extra penalty acts as a disincentive for prey to leave a food-rich location, such as coral reefs, or a response to perceived threats in the outside world. Swirling may be an optimal tactic for prey to remain at the same place while simultaneously evading potential predators.

Figure 8. In a confined space. (a) Without boundary penalty. (b) With boundary penalty.

Figure 9. Effect of speed limit ratio.

4.5. Effect of speed limit ratio

In prior studies, we assume a maximum speed ratio $\|v_0\|_{\max} : \|v_1\|_{\max} = 1 : 1$ between predators and prey. This assumption is reasonable when predator and prey species have similar maximum speeds. Here, we investigate the emergent phenomenon when their maximum speeds have a distinct difference.

As shown in figure 9, adjusting the speed limit ratio to $5 : 3$, we observe more pronounced swarming characteristics evidenced by a smaller DoS dropping from 19% to 18% and a larger DoA rising from 0.82 to around 0.85. Further, tuning the speed limit ratio to $3 : 5$ results in a higher convergence rate of DoS and a lower value of around 17%, attributed to prey's athletic ability to evade predators while still maintaining formation.

4.6. Effect of perception range

In other simulations, predators and prey are assumed to have a perception range equal to the environment size, that is, $R = D$, which is reasonable when organisms have a perception range much larger than their body length. Here, we intentionally tune the perception range to be $R = 2/3D$ and $R = 1/3D$ to investigate its effect on flocking behaviors, as shown in figure 10.

Smaller perception ranges lead to less pronounced flocking behavior, as indicated by a larger DoS and smaller DoA. This effect is especially notable when the perception range is only one-third of the environment size, where DoS increases from 19% to around 20.5%, and DoA drops from 0.82 to 0.75. These findings suggest that perception range plays a crucial role in facilitating the emergence of flocking behaviors.

4.7. Effect of number of predators

In previous simulations, the number of predators present in training is set as $n_0 = 3$. Here, we specifically investigate the special cases where the number of predators during evolution are $n_0 = 0$ and $n_0 = 1$. The episodic evolutions of DoS and DoA are shown in figure 11.

Comparing the cases of a single predator and three predators, we observe that the emergence of swarming behaviors is slightly slower with a single predator. This can be explained by the greater survival pressure exerted by multiple predators, which accelerates the prey's evolution. In the special case *without* predators in the evolution, it can be seen that the DoS and DoA of the prey remain unchanged, indicating that no swarming phenomenon emerges, as exemplified in figure 5(b). This evidence, from another
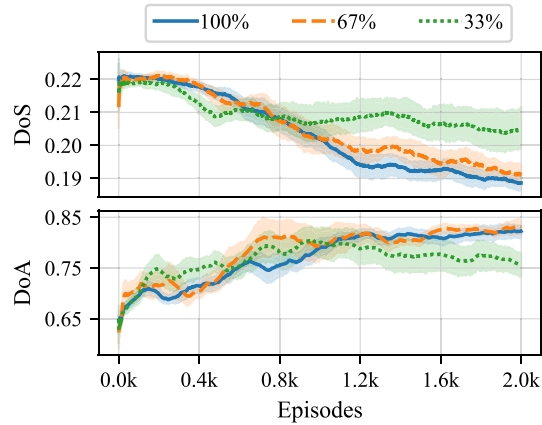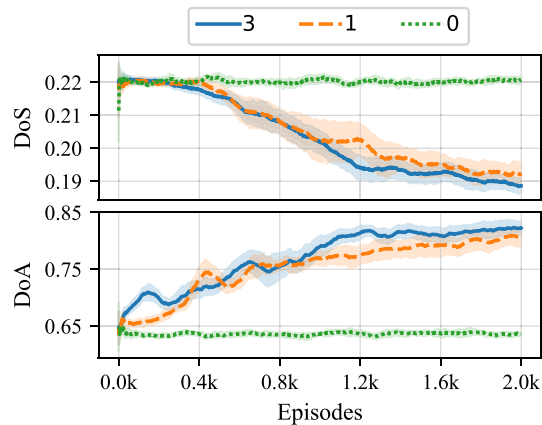
Figure 10. Effect of perception range.

Figure 11. Effect of number of predators.

perspective, further corroborates our hypothesis that predator–prey survival pressure is sufficient to promote swarming behaviors.

5. Discussion and conclusion

In this article, a minimal predator–prey coevolution framework based on mixed cooperative–competitive multiagent RL is proposed, with a swarm-independent reward function based solely on the motivation to survive. We have observed the emergent flocking and swirling behaviors for prey, and dispersion tactic, confusion and marginal predation phenomena for predators. Based on these findings, we hypothesize swarming behaviors can be largely an outcome of passive space extrusion and polarization induced by predators. At the same time, predators may face challenges when trying to select their optimal target from a group of prey. While the proposed framework may not perfectly match the real evolution mechanism in the nature and may not depict the evolution process of all kinds of swarming organisms, it provides a feasible approach for swarming research without the need for handcrafted reward function design.

With the design philosophy of *minimalism*, the proposed framework could serve as a starting point for further experimentation and customization. For instance, one could perform a quantitative analysis on the effect of observation noise, or introduce a third-party species to study emergent behaviors. In the field of robotics, collision penalties can be further incorporated into reward functions to realize collision avoidance simultaneously, enabling robots to exhibit swarming behaviors in a more natural manner without complex handcrafted interaction rules. Overall, the proposed framework and findings could contribute to a better understanding on swarm intelligence of organisms and physical active matter, and have potential applications in swarm robotics.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

Appendix A. Video abstract and animations

The video abstract and simulation animations of figures 2 and 5–8 can be found in the supplementary materials.

Appendix B. Implementation details

We provide more implementation details about the established predator–prey environment and the proposed coevolution algorithm. In the numerical simulation, the agent dynamics described in equation (1) are performed at discrete time steps. For an agent of species i, its states at time step $t+1$ are updated in the following order:

$$\theta(t+1) = \theta(t) + a_\text{R}\Delta t,$$
$$v(t+1) = v(t) + (a_\text{F}h + f_\text{d} + f_\text{a} + f_\text{b})\,\Delta t/m_i,$$
$$x(t+1) = x(t) + v(t)\Delta t,$$

where Δt is the time step with its value shown in table 1. In the periodic boundary condition, when an agent moves beyond one edge of the environment, its position is further updated such that it reappears on the opposite edge as if the edges were connected. For example, if the agent has moved beyond the maximum x-coordinate, we then set its x-coordinate to the minimum x-coordinate plus the distance it has moved beyond the maximum x-coordinate. The agent's actions a_F and a_R are the outputs of the actor neural network, and re-scaled to fit within specified ranges as shown in table 1. Specifically, a_F ranges from zero to its maximum linear acceleration, while a_R ranges from negative maximum angular velocity to positive maximum angular velocity. To calculate f_a, we first analyze whether any two agents collide based on their sizes and distances. If collision occurs, we then apply Hooke's law to calculate the elastic forces resulting from the deformation, and sum them up as $f_\text{a} = \sum_j f_{\text{a},j}$ if multiple collisions happen. The procedure of calculating f_b is the same. The observation vector for an agent has the following form:

$$\begin{bmatrix} \text{agent's own pos., vel. and heading,} \\ \text{relative pos. and headings of observed predators,} \\ \text{relative pos. and headings of observed prey} \end{bmatrix}.$$

The relative positions and headings are reordered from the nearest to the farthest based on range. This is reasonable as conspecifics are considered as homogeneous, as explained in section 3. If the number of observed agents exceeds the topological limit, the farthest ones are removed, and if it does not reach the limit, the rest part of the observation vector is masked out with zeros.

Both critic and actor are encoded by deep feed-forward neural networks with rectified linear unit activation with an input dimension d_o equivalent to the length of the observation vector. Each network consists of three hidden layers with 64 neurons per layer as shown in figure 12, where $d_\text{a} = 2$ is the output dimension of the actor network and a_F, a_R are the output actions.

For a detailed introduction to RL and multi-agent RL, we refer to [18, 39, 40], respectively. For an introduction to multiagent deep deterministic policy gradient algorithm, we refer to [25]. In the training, we first randomly initialize the actor network μ_i parameterized by θ_i^μ and the critic network Q_i parameterized

Table 1. Environment parameters.

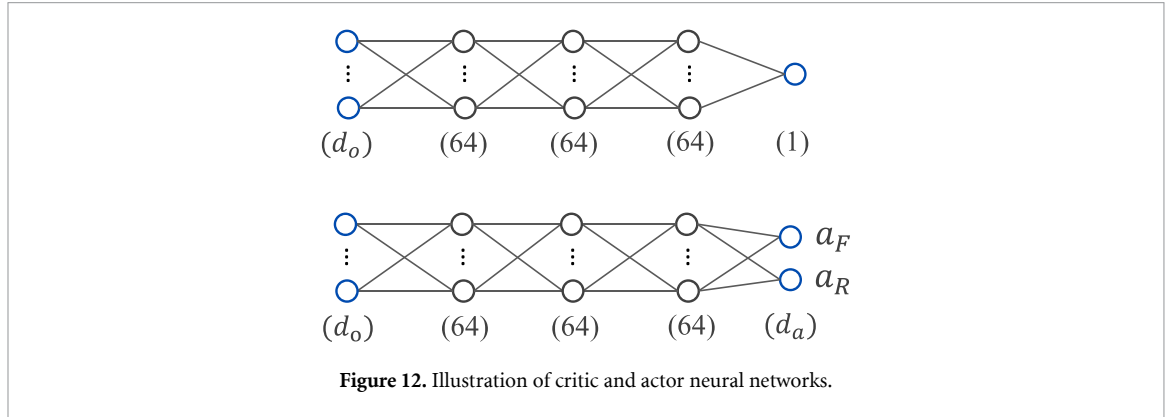| Parameter	Value	Unit
Mass of predator	1	kg
Mass of prey	1	kg
Max speed	0.5 or 0.3	$\mathrm{m\,s^{-1}}$
Max linear acc.	1	$\mathrm{m\,s^{-2}}$
Max angular vel.	0.5	$\mathrm{rad\,s^{-1}}$
Env edge length	2	m
Contact stiffness	50	$\mathrm{N\,m^{-1}}$
Drag coefficient	2	$\mathrm{N{\cdot}s\,m^{-1}}$
Time step	0.1	s

Figure 12. Illustration of critic and actor neural networks.

by θ_i^Q for species i. In this study, predators and prey are denoted as species $i = 0$ and 1, respectively. The corresponding target actor μ_i' and critic Q_i' are also initialized to help mitigate the issue of non-stationarity by slowly updating the value functions as policy improves. At the beginning of each episode, $n_0 \in \mathbb{N}^+$ predators and $n_1 \in \mathbb{N}^+$ prey are placed in the environment at random positions with random headings. Based on current observations $o_i \in \mathbb{R}^{n_i \times d_o}$ where d_o is the dimension of the observation vector, the agents perform actions $a_i \in \mathbb{R}^{n_i \times d_a}$ according to current actors where d_a is the dimension of the action, receive rewards $r_i \in \mathbb{R}^{n_i}$, and obtain new observations o'. The resulting experience tuple (o_i, a_i, r_i, o_i') are saved into replay buffer $\mathcal{B}_i$.

In learning, agents draw a random mini-batch of $S \in \mathbb{N}^+$ samples from $\mathcal{B}_i$, and each sample is denoted as $(o_i^j, a_i^j, r_i^j, o_i'^j) \in (\mathbb{R}^{d_o}, \mathbb{R}^{d_a}, \mathbb{R}, \mathbb{R}^{d_o})$. The critic is updated by minimizing the loss:

$$\mathcal{L}(\theta_i^Q) = \frac{1}{S} \sum_{j=1}^{S} \left(y_i^j - Q_i(o_i^j, a_i^j) \right)^2$$

where the target value $y_i^j = r_i^j + \gamma Q_i'(o_i'^j, a_i'^j)$, $a_i'^j = \mu_i'(o_i'^j)$. To update the actor, the policy gradient theorem [30] is used, and the gradient is approximated as:

$$\nabla_{\theta_i^\mu} J \approx \frac{1}{S} \sum_{j=1}^{S} \nabla_{\theta_i^\mu} \mu_i(o_i^j) \nabla_{\mu_i(o_i^j)} Q_i\left(o_i^j, \mu_i(o_i^j)\right).$$

Finally, the target networks are soft-updated accordingly. The exploration rate ε is the probability that the agent will choose to explore the environment instead of exploiting it. In this study, ε is set to gradually decrease with each episode achieved by using the formula $\max(0.05, \epsilon - 5 \times 10^{-5})$. Similarly, the action noise $\mathcal{N}_t$ also set to decrease gradually with each episode as $\max(0.05, \mathcal{N}_t - 5 \times 10^{-5})$, with their initial values shown in table 2. The learning process is carried out until a state of dynamic equilibrium is achieved between the predators and prey, such that neither party can obtain their future rewards by altering their respective policies. The hyper-parameters for the proposed algorithm are summarized in table 2.

<p align="center">**Table 2.** Hyper-parameters of algorithm.</p>

Hyper-parameter	Value
Number of episodes	2000
Episode length	100
Number of hidden layers	3
Hidden layer size	64
Learning rate of actor	1×10^{-4}
Learning rate of critic	1×10^{-3}
Discount factor	0.95
Soft-update rate	0.01
Initial exploration rate	0.1
Initial noise rate	0.1
Replay buffer size	5×10^5
Batch size	256

ORCID iDs

Jianan Li ● https://orcid.org/0000-0002-9733-3244
Liang Li ● https://orcid.org/0000-0002-2447-6295
Shiyu Zhao ● https://orcid.org/0000-0003-3098-8059

References

[1] Sumpter D J 2010 Collective animal behavior *Collective Animal Behavior* (Princeton University Press)
[2] Krause J, Krause P, Ruxton G and Ruxton G 2002 *Living in Groups* (*Oxford Series in Ecology and Evolution*) (Oxford University Press)
[3] Vicsek T, Czirók A, Ben-Jacob E, Cohen I and Shochet O 1995 Novel type of phase transition in a system of self-driven particles *Phys. Rev. Lett.* **75** 1226
[4] Chung S-J, Paranjape A A, Dames P, Shen S and Kumar V 2018 A survey on aerial swarm robotics *IEEE Trans. Robot.* **34** 837
[5] Couzin I D, Krause J, James R, Ruxton G D and Franks N R 2002 Collective memory and spatial sorting in animal groups *J. Theor. Biol.* **218** 1
[6] Jia Y and Vicsek T 2019 Modelling hierarchical flocking *New J. Phys.* **21** 093048
[7] Pearce D J, Miller A M, Rowlands G and Turner M S 2014 Role of projection in the control of bird flocks *Proc. Natl Acad. Sci.* **111** 10422
[8] Barberis L and Peruani F 2016 Large-scale patterns in a minimal cognitive flocking model: incidental leaders, nematic patterns and aggregates *Phys. Rev. Lett.* **117** 248001
[9] Lavergne F A, Wendehenne H, Bäuerle T and Bechinger C 2019 Group formation and cohesion of active particles with visual perception–dependent motility *Science* **364** 70
[10] Bastien R and Romanczuk P 2020 A model of collective behavior based purely on vision *Sci. Adv.* **6** eaay0792
[11] Chen Y and Kolokolnikov T 2014 A minimal model of predator–swarm interactions *J. R. Soc. Interface* **11** 20131208
[12] Liu D, Liang Y, Deng J and Zhang W 2023 Modeling three-dimensional bait ball collective motion *Phys. Rev.* E **107** 014606
[13] Olson R S, Hintze A, Dyer F C, Knoester D B and Adami C 2013 Predator confusion is sufficient to evolve swarming behaviour *J. R. Soc. Interface* **10** 20130305
[14] Sunehag P, Lever G, Liu S, Merel J, Heess N, Leibo J Z, Hughes E, Eccles T and Graepel T 2019 Reinforcement learning agents acquire flocking and symbiotic behaviour in simulated ecosystems *ALIFE 2019: The 2019 Conf. on Artificial Life* (MIT Press) pp 103–10
[15] Hahn C, Phan T, Gabor T, Belzner L and Linnhoff-Popien C 2019 Emergent escape-based flocking behavior using multi-agent reinforcement learning *ALIFE 2019: The 2019 Conf. on Artificial Life* (MIT Press) pp 598–605
[16] Durve M, Peruani F and Celani A 2020 Learning to flock through reinforcement *Phys. Rev.* E **102** 012601
[17] Monter S, Heuthe V-L, Panizon E and Bechinger C 2023 Dynamics and risk sharing in groups of selfish individuals *J. Theor. Biol.* **562** 111433
[18] Sutton R S and Barto A G 2018 *Reinforcement Learning: An Introduction* (MIT Press)
[19] Mui nos-Landin S, Fischer A, Holubec V and Cichos F 2021 Reinforcement learning with artificial microswimmers *Sci. Robot.* **6** eabd9285
[20] Nasiri M and Liebchen B 2022 Reinforcement learning of optimal active particle navigation *New J. Phys.* **24** 073042
[21] Kaiser T K and Hamann H 2022 Innate motivation for robot swarms by minimizing surprise: from simple simulations to real-world experiments *IEEE Trans. Robot.* **38** 3582
[22] Nelson A L, Barlow G J and Doitsidis L 2009 Fitness functions in evolutionary robotics: a survey and analysis *Robot. Auton. Syst.* **57** 345
[23] Hamilton W D 1971 Geometry for the selfish herd *J. Theor. Biol.* **31** 295
[24] Mordatch I and Abbeel P 2017 Emergence of grounded compositional language in multi-agent populations (arXiv:1703.04908)
[25] Lowe R, Wu Y, Tamar A, Harb J, Abbeel P and Mordatch I 2017 Multi-agent actor-critic for mixed cooperative-competitive environments *31st Conf. on Neural Information Processing Systems (NIPS)* pp 6382–93
[26] Frenkel D and Smit B 2001 *Understanding Molecular Simulation: From Algorithms to Applications* vol 1 (Elsevier)
[27] Ballerini M *et al* 2008 Interaction ruling animal collective behavior depends on topological rather than metric distance: evidence from a field study *Proc. Natl Acad. Sci.* **105** 1232

[28] Moussaïd M, Helbing D and Theraulaz G 2011 How simple rules determine pedestrian behavior and crowd disasters *Proc. Natl Acad. Sci.* **108** 6884

[29] Hüttenrauch M *et al* 2019 Deep reinforcement learning for swarm systems *J. Mach. Learn. Res.* **20** 1966–96

[30] Sutton R S, McAllester D, Singh S and Mansour Y 1999 Policy gradient methods for reinforcement learning with function approximation *Advances in Neural Information Processing Systems* vol 12 pp 1057–63

[31] Reynolds C W 1987 Flocks, herds and schools: a distributed behavioral model *Proc. 14th Annual Conf. on Computer Graphics and Interactive Techniques* vol 21 pp 25–34

[32] Huffaker C *et al* 1958 Experimental studies on predation: dispersion factors and predator-prey oscillations *Hilgardia* **27** 343

[33] Major P F 1978 Predator-prey interactions in two schooling fishes, Caranx ignobilis and Stolephorus purpureus *Animal Behav.* **26** 760

[34] Duffield C and Ioannou C C 2017 Marginal predation: do encounter or confusion effects explain the targeting of prey group edges? *Behav. Ecol.* **28** 1283

[35] Romenskyy M, Herbert-Read J E, Ioannou C C, Szorkovszky A, Ward A J and Sumpter D J 2020 Quantifying the structure and dynamics of fish shoals under predation threat in three dimensions *Behav. Ecol.* **31** 311

[36] Krakauer D C 1995 Groups confuse predators by exploiting perceptual bottlenecks: a connectionist model of the confusion effect *Behav. Ecol. Sociobiol.* **36** 421

[37] Parrish J K and Edelstein-Keshet L 1999 Complexity, pattern and evolutionary trade-offs in animal aggregation *Science* **284** 99–101

[38] Franks N R, Worley A, Grant K A, Gorman A R, Vizard V, Plackett H, Doran C, Gamble M L, Stumpe M C and Sendova-Franks A B 2016 Social behaviour and collective motion in plant-animal worms *Proc. R. Soc.* B **283** 20152946

[39] Zhao S 2024 *Mathematical Foundations of Reinforcement Learning* (Tsinghua University Press and Springer)

[40] Zhang K, Yang Z and Başar T 2021 Multi-agent reinforcement learning: a selective overview of theories and algorithms *Handbook of Reinforcement Learning and Control* (Springer) pp 321–84